Alzheimer's Disease Analysis Using Regression Models

Dennis Feng, Hongbo Yin, Soyoung Chung and Zhiyi Da

Dec. 2, 2021

Report Contribution

- **Soyoung Chung and Hongbo Yin**:
  - Contributed to the introduction
  - Contributed to the Two Sample T-test Part
- **Zhiyi Da**:
  - Contributed to the Multiple Regression Analysis Part
- **Dennis Feng**:
  - Contributed to the Effect of missing data Part

Abstract

In this project, our group attempts to estimate the various factors of Alzheimer's disease using R. The objective of this project is to design a linear regression model and set hypothesis testing in order to check which independent variables are significant or not. For our first hypothesis, we are interested in finding out whether there is a relationship between gender and Alzheimer's disease. For our second hypothesis, we are interested in fitting a multiple linear regression model in order to find the most significant independent variables for Alzheimer's disease.

Installation

```
install.packages(c("mice","VIM"))
install.packages("leaps")
install.packages("Rcpp")
install.packages("lattice")
install.packages("parallel")
```

```
library(mice)
library(VIM)
library(leaps)
library(simFrame)
library(Rcpp)
library(lattice)
```

Introduction

According to the Alzheimer's Association, Alzheimer's disease is a type of disease that negatively affects memory, thinking and behavior. Oftentimes, Alzheimer's disease spirals out of control with the people who have it and it severely affects their life especially day to day activities. There has yet to be a cure for this disease but there is a lot of research going on pertaining to Alzheimer's disease. There are also a lot of

unknown causes pertaining to Alzheimer's disease. The dataset that we used is a cross-sectional dataset found on kaggle.com and the dataset is from OASIS (Open Access Series of Imaging Studies). There are 436 rows of 12 variables

ID Identification(436 unique values)

M/F gender

Hand Dominant Hand (1 unique values)

Age Age in years

Educ Education Level

SES Socioeconomic Status

MMSE Mini Mental State Examination

CDR Clinical Dementia Rating

eTIV Estimated Total Intracranial Volume

nWBV Normalize Whole Brain Volume

ASF Atlas Scaling Factor.

There are also some missing values and this will be addressed further below. We will use nWBV as our measure for Alzheimer's disease. The more nWBV that one person has, the less likely that person is going to have Alzheimer's disease. The less nWBV that one person has, the more likely that person is going to have Alzheimer's disease.

Two Sample T-test

First hypothesis

For the first hypothesis, we are interested in finding out whether there is a relationship between gender and Alzheimer's disease. Let's say μ_f be the average nWBV of female and μ_m be the average nWBV of male.

H_0: There is no linear association between Dementia and Gender = There is no significant difference between Dementia of male and female.

H_1: There is a linear association between Dementia and Gender. = There is a significant difference between Dementia of male and female.

H_0: $\mu_f = \mu_m$ vs. H_1: $\mu_f \neq \mu_m$

Methodology

1. Load and convert data

We use R to read 'oasis_cross-sectional.csv' files for data analysis.

```
setwd("/Users/dazhiyi/Downloads/sbu/2021/ams572_project/")
data = read.csv('oasis_cross-sectional.csv')
```

Since the gender data has character 'F' of female and 'M'of male, we need to change each of characters to numbers. Female becomes 1, male becomes 2.

```
data$M.F=as.factor(data$M.F)
#Female=1; Male =2;
str(data$M.F)
```

```
##  Factor w/ 2 levels "F","M": 1 1 1 2 2 1 2 1 2 1 ...
```

2. Extract necessary variables from data

Make an matrix that only has gender and nWBV variables.

```
mymatrix=matrix(c(data$M.F,data$nWBV),ncol = 2,byrow = F)
```

Let nWBV_F be a matrix that only has female and nWBV variables.

Let nWBV_M be a matrix that only has male and nWBV variables.

```
nWBV_F=mymatrix[which(mymatrix[,1]==1),]
nWBV_M=mymatrix[which(mymatrix[,1]==2),]
```

3. Use the two-sample t-test

The two-sample t-test is a method that tests whether the unknown population means of two groups are equal or not. We use two-sample t-test to test whether average nWBV of female and average nWBV of male are equal or not. We can use the two-sample t-test when our data values are independent and the two independent groups have equal variances. We first need to check if nWBV_F and nWBV_M have same variance

Check if two groups have equal variance

```
var.test(nWBV_F[,2],nWBV_M[,2])
```

```
##
##  F test to compare two variances
##
## data:  nWBV_F[, 2] and nWBV_M[, 2]
## F = 0.97052, num df = 267, denom df = 167, p-value = 0.8218
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.7345048 1.2711736
## sample estimates:
## ratio of variances
##          0.9705219
```

Since the p-value is greater than significant level of 0.05, we fail to reject the null hypothesis. Then we can say the variances of two groups are equal.

Two Sample t-test

Since the variance of two groups are equal, we can use the two-sample t-test with **var.equal = TRUE**.

```
res<-t.test(nWBV_F[,2],nWBV_M[,2], var.equal = TRUE)
res
```

```
##
##  Two Sample t-test
##
## data:  nWBV_F[, 2] and nWBV_M[, 2]
## t = 0.05988, df = 434, p-value = 0.9523
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.01125221  0.01195939
## sample estimates:
## mean of x mean of y
## 0.7918060 0.7914524
```

```
res$p.value
```

```
## [1] 0.9522784
```

The p-value is bigger than 0.05, which means we do not reject H_0.

Conclusion for first hypothesis
Therefore, we can conclude that there is insufficient evidence showing that there is significant difference between Dementia of male and female

4. Effect of missing data

```
mydata <- read.csv("oasis_cross-sectional.csv")
mydataFrame<-as.data.frame(mydata)
nac<-NAControl(NArate=0.2)
x<-setNA(mydataFrame,nac)
M.F_MAR=as.factor(x$M.F)
mymatrix_MAR=matrix(c(M.F_MAR,x$nWBV),ncol = 2,byrow = F)
```

1. Missing Values at Random(MCAR) Delete missing values

```
newdata=na.omit(mymatrix_MAR)
nWBV_F_MAR=newdata[which(newdata[,1]==1),]
nWBV_M_MAR=newdata[which(newdata[,1]==2),]
```

Check equal variance assumption

```
var.test(nWBV_F_MAR[,2],nWBV_M_MAR[,2])
```

```
##
##  F test to compare two variances
##
## data:  nWBV_F_MAR[, 2] and nWBV_M_MAR[, 2]
## F = 0.91382, num df = 180, denom df = 97, p-value = 0.5996
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.6374602 1.2857740
## sample estimates:
## ratio of variances
##          0.9138186
```

Since the p-value is bigger than 0.05 which means equal variance.

Two sample t-test with equal variance

```
t.test(nWBV_F_MAR[,2],nWBV_M_MAR[,2],var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  nWBV_F_MAR[, 2] and nWBV_M_MAR[, 2]
## t = -0.01564, df = 277, p-value = 0.9875
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.01533410  0.01509236
## sample estimates:
## mean of x mean of y
## 0.7902873 0.7904082
```

Conclusion

The p-value is bigger than 0.05. DO NOT Reject H_0. There is insufficient evidence showing that there is significant difference between Dementia of male and female.

2. Compare the data with missing values at random and original data If we look at the mean x and mean y with and without missing values at randomly, we can see that the difference is very samll, so there is not a significant difference.

3. Missing Values Not At Random(MNAR) Delete missing values

```
nWBV_F_MNAR=replace(nWBV_F[,2],1:100,NA)
```

Replace first 100 values in nWBV of female to be missing values

```
new_nWBV_F_MNAR=na.omit(nWBV_F_MNAR)
```

Check equal variance assumption

```
var.test(new_nWBV_F_MNAR,nWBV_M[,2])
```

```
##
##  F test to compare two variances
##
## data:  new_nWBV_F_MNAR and nWBV_M[, 2]
## F = 0.95156, num df = 167, denom df = 167, p-value = 0.7487
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.7018602 1.2901018
## sample estimates:
## ratio of variances
##          0.9515624
```

The p-value is bigger than 0.05 which means equal variance.

Two sample t-test with equal variance

```
t.test(new_nWBV_F_MNAR,nWBV_M_MAR[,2],var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  new_nWBV_F_MNAR and nWBV_M_MAR[, 2]
## t = 0.068985, df = 264, p-value = 0.9451
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.01466102  0.01572565
## sample estimates:
## mean of x mean of y
## 0.7909405 0.7904082
```

The p-value is bigger than 0.05. DO NOT Reject H_0. There is insufficient evidence showing that there is significant difference between Dementia of male and female.

Conclusion

Compare the data with missing values not at random and original data: If we look at the mean x and mean y with and without missing values not at randomly, we can see that the difference is still very small, so there is not a significant difference.

Multiple Regression Analysis

Second hypothesis

For the second hypothesis, we are interested in finding out which variable is significant. Our null hypothesis is that each variable is equal to 0 and our alternative hypothesis is that at least one variable is not equal to 0, making that variable significant.

H_0: $\mu_{M.FM} = \mu_{Age} = \mu_{Educ} = \mu_{SES} = \mu_{MMSE} = \mu_{CDR} = \mu_{eTIV} = \mu_{ASF}$

H_1: $\mu_{M.FM} \neq \mu_{Age} \neq \mu_{Educ} \neq \mu_{SES} \neq \mu_{MMSE} \neq \mu_{CDR} \neq \mu_{eTIV} \neq \mu_{ASF}$

Method steps

We use R to read 'oasis_cross-sectional.csv' files for data analysis. Through using **stringsAsFactors = TRUE**, it is appropriate to convert nominal variables to factor variables.

```
setwd("/Users/dazhiyi/Downloads/sbu/2021/ams572_project/")
data = read.csv('oasis_cross-sectional.csv',stringsAsFactors = TRUE)
# Verify that the data is converted to the form we expected earlier
str(data)
```

```
## 'data.frame':    436 obs. of  12 variables:
##  $ ID   : Factor w/ 436 levels "OAS1_0001_MR1",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ M.F  : Factor w/ 2 levels "F","M": 1 1 1 2 2 1 2 1 2 1 ...
##  $ Hand : Factor w/ 1 level "R": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Age  : int  74 55 73 28 18 24 21 20 74 52 ...
##  $ Educ : int  2 4 4 NA NA NA NA NA 5 3 ...
##  $ SES  : int  3 1 3 NA NA NA NA NA 2 2 ...
##  $ MMSE : int  29 29 27 NA NA NA NA NA 30 30 ...
##  $ CDR  : num  0 0 0.5 NA NA NA NA NA 0 0 ...
##  $ eTIV : int  1344 1147 1454 1588 1737 1131 1516 1505 1636 1321 ...
##  $ nWBV : num  0.743 0.81 0.708 0.803 0.848 0.862 0.83 0.843 0.689 0.827 ...
##  $ ASF  : num  1.31 1.53 1.21 1.1 1.01 ...
##  $ Delay: Factor w/ 15 levels "1","10","12",..: 15 15 15 15 15 15 15 15 15 15 ...
```

1. Analyze whether missing value exists in data.

```
# sum of the rows with one or more missing values
sum(!complete.cases(data))
```

```
## [1] 220
```

```
mean(is.na(data)) #15.7%of instances have missing values
```

```
## [1] 0.1573012
```

```
mean(!complete.cases(data))#50% of the instances contained one or more missing values
```

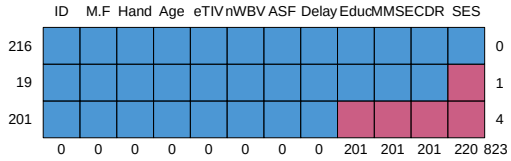
```
## [1] 0.5045872
```

We can use the **md.pattern()** function in **mice** package to observe the missing values more intuitively, which it generates a table showing the pattern of missing values in the form of a matrix or data box.

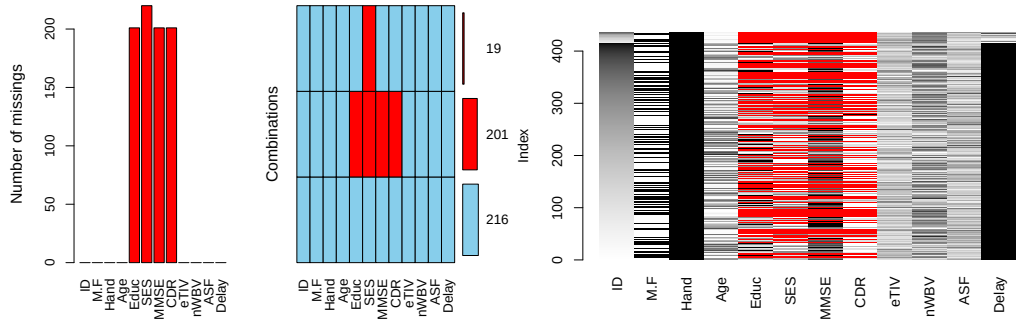
```
md.pattern(data)
```

```
##     ID M.F Hand Age eTIV nWBV ASF Delay Educ MMSE CDR SES
## 216  1   1    1   1    1    1   1     1    1    1   1   1   0
## 19   1   1    1   1    1    1   1     1    1    1   1   0   1
```

```
## 201   1   1    1   1    1    1   1     1    0    0   0   0   4
##       0   0    0   0    0    0   0     0  201  201 201 220 823
```




```
#aggr () function generates a pattern of missing values for the dataset
aggr(data,prop=F,numbers=T)
matrixplot(data)
```



2. To use Multiple interpolation to impute data.

Through looking "mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software" paper, We learned how to use mice to impute data.

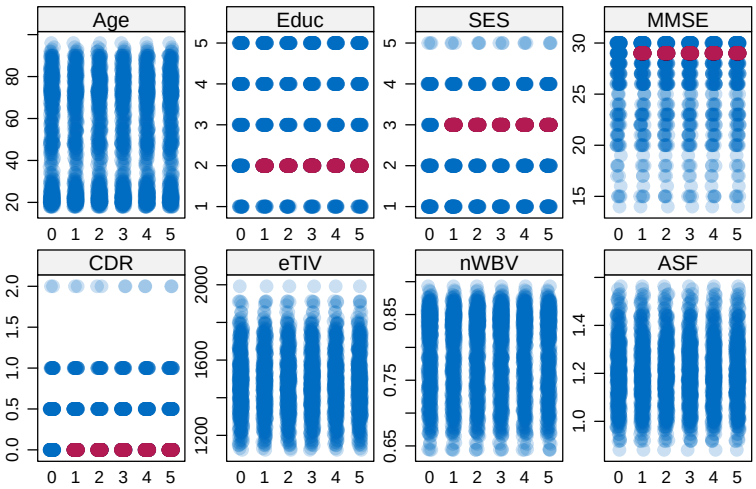
picture reference on "mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software"

```
imp = mice(data,seed = 1234)
```

```
ok = with(imp,lm(nWBV~M.F+Age+Educ+SES+MMSE+CDR+eTIV+ASF,data=data))
pooled = pool(ok)
summary(pooled)
```

```
##          term      estimate    std.error    statistic      df    p.value
## 1 (Intercept)  7.241255e-01 2.559889e-01   2.82873778 205.008 0.00513723
## 2        M.FM -6.040661e-03 5.069491e-03  -1.19157137 205.008 0.23480698
## 3         Age -2.506020e-03 1.707489e-04 -14.67663642 205.008 0.00000000
## 4        Educ -4.402378e-05 2.245368e-03  -0.01960649 205.008 0.98437636
## 5         SES  1.333788e-03 2.625522e-03   0.50800853 205.008 0.61199352
## 6        MMSE  2.224686e-03 9.285244e-04   2.39593723 205.008 0.01747675
## 7         CDR -2.037820e-02 8.279082e-03  -2.46140754 205.008 0.01466544
## 8        eTIV  3.048820e-05 8.686763e-05   0.35097304 205.008 0.72596898
## 9         ASF  8.774335e-02 1.068240e-01   0.82138238 205.008 0.41238267
```

```
data2 = complete(imp,action = 3)
stripplot(imp,pch=19,cex=1.2,alpha=.3)
```



3. Building model

We will use the multiple linear regression to build the model. Multiple linear regression is an extension of simple linear regression model. It can evaluate more complex relationships because it can predict a response variable (y) based on multiple different predictor variables (x). In multiple regression we fit a model of the form(excluding the error)

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$$

where $x_1, x_2, ..., x_k$ are $k \geq 2$ predictor variables and $\beta_0, \beta_1, \beta_2, ..., \beta_k$ are $k+1$ unknown parameters. check how goodness of fit this model: we use the residuals defined by

$$\varepsilon_i = y_i - \hat{y}_i \qquad (i = 1, 2, ..., n)$$

where the $\bar{y}_i$ are the fitted values:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + ... + \hat{\beta}_k x_{ik} \qquad (i = 1, 2, ..., n)$$

we can use Chapter 10 error sum of squares as overall measure:

$$SSE = \sum_{i=1}^{n} \varepsilon_i^2$$

we compare the SSE to the total sum of squares, $SST = \sum (y_i - \bar{y})^2$. As in Chapter 10, define the regression sum of squares given by:

$$SSR = SST - SSE$$

The coefficient of multiple determination is the ratio of SSR to SST:

$$r^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

In multiple linear regression, the r^2 represents the correlation coefficient between the observed values of the response variable (y) and the fitted values of y. Thus, the value of r will always be positive and will range from 0 to 1, with values closer to 1 representing better fits. The higher r^2, the better the model. We want to build a model for estimating Normalize Whole Brain Volume based on Gender, Age, Education Level, Socioeconomic Status, Mini Mental State Examination, Clinical Dementia Rating, Estimated Total Intracranial Volume, and Atlas Scaling Factor, as follows:

$$nWBV = \beta_0 + \beta_1 M.F + \beta_2 Age + \beta_3 Educ + \beta_4 SES + \beta_5 MMSE + \beta_6 CDR + \beta_7 eTIV + \beta_8 ASF$$

we can compute the model's coefficients in R:

```r
fit = lm(nWBV~M.F+Age+Educ+SES+MMSE+CDR+eTIV+ASF,data = data2)
summary(fit)
```

```
##
## Call:
## lm(formula = nWBV ~ M.F + Age + Educ + SES + MMSE + CDR + eTIV +
##     ASF, data = data2)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.083928 -0.015437  0.000996  0.017298  0.069807
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.183e-01  1.052e-01   7.780 5.52e-14 ***
## M.FM        -2.526e-03  2.996e-03  -0.843   0.3996
## Age         -1.890e-03  6.345e-05 -29.785  < 2e-16 ***
## Educ         2.151e-03  1.780e-03   1.208   0.2277
## SES          2.168e-03  2.258e-03   0.960   0.3375
## MMSE         2.717e-03  6.877e-04   3.950 9.12e-05 ***
## CDR         -2.118e-02  6.628e-03  -3.195   0.0015 **
## eTIV        -2.627e-05  3.509e-05  -0.749   0.4545
## ASF          2.167e-02  4.312e-02   0.503   0.6156
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02529 on 427 degrees of freedom
## Multiple R-squared:  0.8253, Adjusted R-squared:  0.822
## F-statistic: 252.1 on 8 and 427 DF,  p-value: < 2.2e-16
```

After checking the summary, we can find the following results:

$Intercept$	$M.F = x_1$	$Age = x_2$	$Educ = x_3$	$SES = x_4$	$MMSE = x_5$	$CDR = x_6$	$eTIV = x_7$	$ASF = x_8$
$8.183e^{-1}$	$-2.526e^{-3}$	$-1.890e^{-3}$	$2.151e^{-3}$	$2.168e^{-3}$	$2.717e^{-3}$	$-2.118e^{-2}$	$-2.627e^{-5}$	$2.167e^{-2}$

The model equation can be written as follows:

$$nWBV = 8.183e^{-1} - 2.526e^{-3}x_1 - 1.890e^{-3}x_2 + 2.151e^{-3}x_3 + 2.168e^{-3}x_4 + 2.717e^{-3}x_5 - 2.118e^{-2}x_6 - 2.627e^{-5}x_7 + 2.167e^{-2}x_8$$

4. Use Best subsets regression to find the best model.

Best subset regression is a model selection method that tests all possible combinations of predictor variables. We can determine which are the best predictor variables by using best subset regression, and then confirm the optimal model according to certain statistical criteria.

Through using **regsubsets()** function in **leaps** package, we can find the best predictor variables in the linear model.

```r
leap = regsubsets(nWBV~M.F+Age+Educ+SES+MMSE+CDR+eTIV+ASF,data = data2)
summary(leap)
```

```
## Subset selection object
## Call: regsubsets.formula(nWBV ~ M.F + Age + Educ + SES + MMSE + CDR +
##     eTIV + ASF, data = data2)
```

```
## 8 Variables  (and intercept)
##      Forced in Forced out
## M.FM     FALSE      FALSE
## Age      FALSE      FALSE
## Educ     FALSE      FALSE
## SES      FALSE      FALSE
## MMSE     FALSE      FALSE
## CDR      FALSE      FALSE
## eTIV     FALSE      FALSE
## ASF      FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: exhaustive
##          M.FM Age Educ SES MMSE CDR eTIV ASF
## 1  ( 1 ) " "  "*" " "  " " " "  " " " "  " "
## 2  ( 1 ) " "  "*" " "  " " " "  "*" " "  " "
## 3  ( 1 ) " "  "*" " "  " " "*"  " " "*"  " "
## 4  ( 1 ) " "  "*" " "  " " "*"  "*" "*"  " "
## 5  ( 1 ) "*"  "*" " "  " " "*"  "*" "*"  " "
## 6  ( 1 ) " "  "*" "*"  "*" "*"  "*" "*"  " "
## 7  ( 1 ) "*"  "*" "*"  "*" "*"  "*" "*"  " "
## 8  ( 1 ) "*"  "*" "*"  "*" "*"  "*" "*"  "*"
```

Through **summary()** function, we get some optimality criteria. We will use r^2,*adjuested* r^2,C_p,BIC to select best model.

```
leap_s = summary(leap)
names(leap_s)
```

```
## [1] "which"  "rsq"    "rss"    "adjr2"  "cp"     "bic"    "outmat" "obj"
```

We find that the r^2 statistic increases from 76.4% to 82.5%. Thus, the r^2 statistic increases monotonically as more variables are included.

```
leap_s$rsq
```

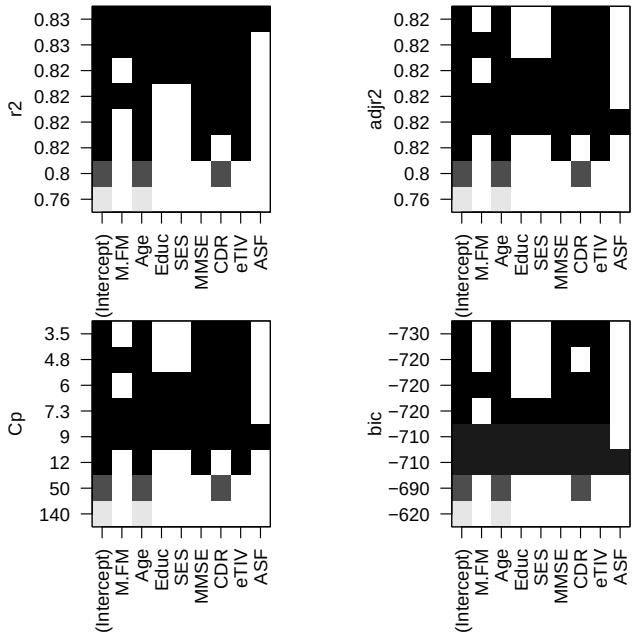
```
## [1] 0.7640509 0.8034973 0.8199584 0.8242737 0.8245558 0.8248624 0.8251735
## [8] 0.8252769
```

We can see that the following model selection have the same number 4 for the best set of predictor variables.

```
data.frame(
  adj_r2_max = which.max(leap_s$adjr2),
  cp_min = which.min(leap_s$cp),
  bic_min = which.min(leap_s$bic)
)
```

```
##   adj_r2_max cp_min bic_min
## 1          4      4       4
```

```
par(mfrow = c(1,2))
plot(leap, scale = "r2")
plot(leap, scale = "adjr2")
par(mfrow = c(1,2))
plot(leap, scale = "Cp")
plot(leap, scale = "bic")
```



Finally, we get the best 4 predictor variables to make the best linear model.

```
coef(leap, 4)
```

```
##   (Intercept)           Age          MMSE           CDR          eTIV
##  8.840233e-01 -1.860291e-03  2.763700e-03 -2.149933e-02 -4.775729e-05
```

The model equation can be written as follows:

$Intercept$	$Age = x_2$	$MMSE = x_5$	$CDR = x_6$	$eTIV = x_7$
$8.8402e^{-1}$	$-1.8602e^{-3}$	$2.7637e^{-3}$	$-2.1499e^{-2}$	$-4.7757e^{-5}$

$$nWBV = 8.8402e^{-1} - 1.8602e^{-3}x_2 + 2.7637^{-3}x_5 - 2.1499e^{-2}x_6 - 4.7757e^{-5}x_7$$

```
f = lm(nWBV~Age+MMSE+CDR+eTIV,data=data2)
summary(f)
```

```
##
## Call:
## lm(formula = nWBV ~ Age + MMSE + CDR + eTIV, data = data2)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.083061 -0.015774  0.000805  0.017497  0.073027
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.840e-01  2.232e-02  39.608  < 2e-16 ***
## Age         -1.860e-03  5.606e-05 -33.183  < 2e-16 ***
## MMSE         2.764e-03  6.767e-04   4.084 5.28e-05 ***
## CDR         -2.150e-02  6.608e-03  -3.253  0.00123 **
## eTIV        -4.776e-05  7.752e-06  -6.161 1.66e-09 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02524 on 431 degrees of freedom
## Multiple R-squared:  0.8243, Adjusted R-squared:  0.8226
## F-statistic: 505.4 on 4 and 431 DF,  p-value: < 2.2e-16
```

We can see that these 4 predictor variables give a better r^2 value which leads to a better and more concise linear model than a model with 8 predictor variables. Thus, these 4 predictor variables will have the best linear model, as following,

$$nWBV = 8.8402e^{-1} - 1.8602e^{-3}Age + 2.7637^{-3}MMSE - 2.1499e^{-2}CDR - 4.7757e^{-5}eTIV$$

5. Effect of missing data.

We believe that our data has missing not at random values. This is likely due to the fact that the patient may have refused testing on a certain test, or refused to answer some questions during the questionnaire. For example, the questionnaire asks about education level and social status. This may not have been a mandatory question needed to be answered or the patient may not feel like answering this question due to social norms or perhaps embarrassment. Regardless, this creates a clear bias in the data and thus the analysis. For the missing not at random values, we have decided to use MICE to impute the missing values from the data set. The hypothesis that we are testing is the same as hypothesis 2.

This is the model without imputing for missing data

```
fit1=lm(nWBV~M.F+Age+Educ+SES+MMSE+CDR+eTIV+ASF,data=data)
summary(fit1)
```

```
##
## Call:
## lm(formula = nWBV ~ M.F + Age + Educ + SES + MMSE + CDR + eTIV +
##     ASF, data = data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.077372 -0.018709  0.000512  0.018642  0.073261
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.241e-01  2.560e-01   2.829  0.00513 **
## M.FM        -6.041e-03  5.069e-03  -1.192  0.23479
## Age         -2.506e-03  1.708e-04 -14.677  < 2e-16 ***
## Educ        -4.402e-05  2.245e-03  -0.020  0.98438
## SES          1.334e-03  2.626e-03   0.508  0.61199
## MMSE         2.225e-03  9.285e-04   2.396  0.01747 *
## CDR         -2.038e-02  8.279e-03  -2.461  0.01466 *
## eTIV         3.049e-05  8.687e-05   0.351  0.72597
## ASF          8.774e-02  1.068e-01   0.821  0.41237
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02866 on 207 degrees of freedom
##   (220 observations deleted due to missingness)
## Multiple R-squared:  0.6605, Adjusted R-squared:  0.6474
## F-statistic: 50.35 on 8 and 207 DF,  p-value: < 2.2e-16
```

$$nWBV = 7.241e^{-1} - 6.041e^{-3}x_1 - 2.506e^{-3}x_2 - 4.402e^{-4}x_3 + 1.334e^{-3}x_4 + 2.225e^{-3}x_5 - 2.038e^{-2}x_6 + 3.049e^{-5}x_7 + 8.774e^{-2}x_8$$

This is the model when imputations are done for the missing data from method step 2:

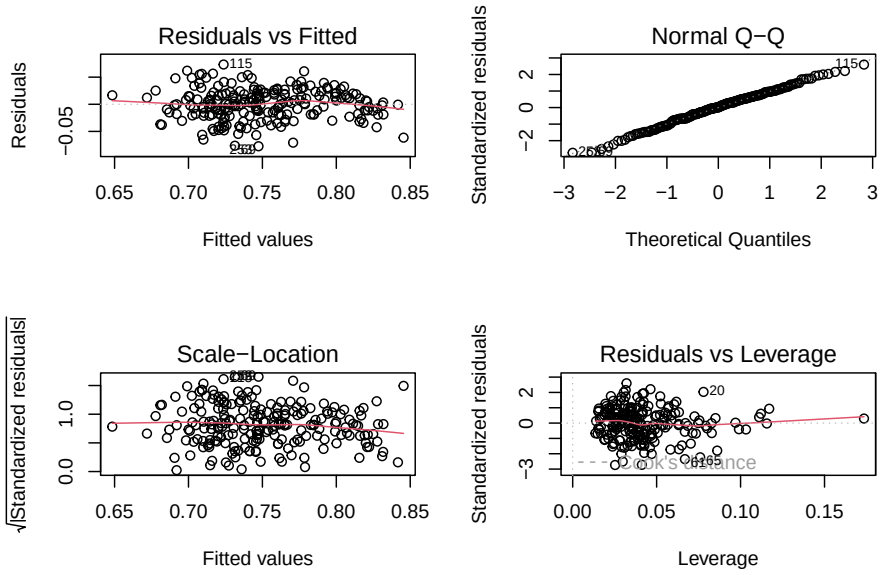
```
fit2 = lm(nWBV~M.F+Age+Educ+SES+MMSE+CDR+eTIV+ASF,data = data2)
summary(fit2)
```

```
##
## Call:
## lm(formula = nWBV ~ M.F + Age + Educ + SES + MMSE + CDR + eTIV +
##     ASF, data = data2)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.083928 -0.015437  0.000996  0.017298  0.069807
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.183e-01  1.052e-01   7.780 5.52e-14 ***
## M.FM        -2.526e-03  2.996e-03  -0.843   0.3996
## Age         -1.890e-03  6.345e-05 -29.785  < 2e-16 ***
## Educ         2.151e-03  1.780e-03   1.208   0.2277
## SES          2.168e-03  2.258e-03   0.960   0.3375
## MMSE         2.717e-03  6.877e-04   3.950 9.12e-05 ***
## CDR         -2.118e-02  6.628e-03  -3.195   0.0015 **
## eTIV        -2.627e-05  3.509e-05  -0.749   0.4545
## ASF          2.167e-02  4.312e-02   0.503   0.6156
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02529 on 427 degrees of freedom
## Multiple R-squared:  0.8253, Adjusted R-squared:  0.822
## F-statistic: 252.1 on 8 and 427 DF,  p-value: < 2.2e-16
```

$$nWBV = 8.183e^{-1} - 2.526e^{-3}x_1 - 1.890e^{-3}x_2 + 2.151e^{-4}x_3 + 2.168e^{-3}x_4 + 2.717e^{-3}x_5 - 2.118e^{-2}x_6 - 2.627e^{-5}x_7 + 2.167e^{-2}x_8$$

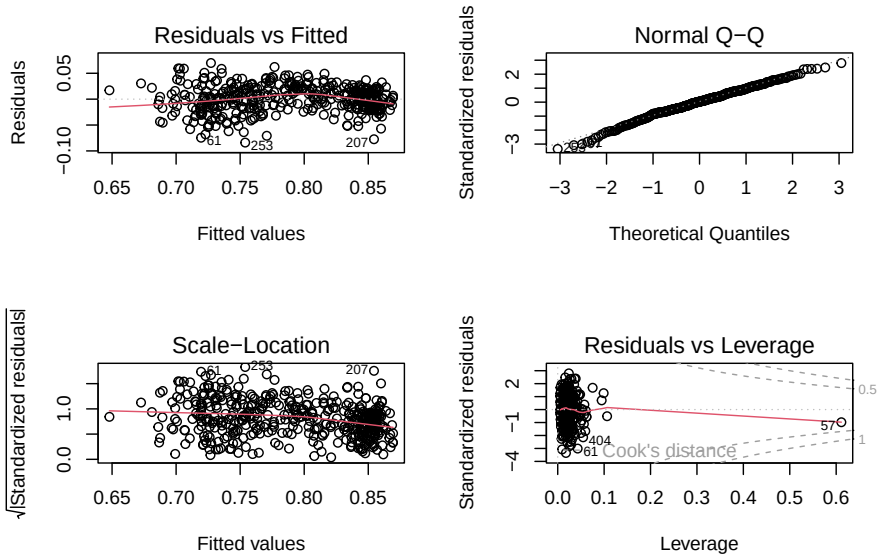
Below are plots of the first model

```
par(mfrow = c(2,2))
plot(fit1)
```



Below are plots of the second model

```
par(mfrow = c(2,2))
plot(fit2)
```



As you can see from both outputs, the fit (r^2 value) for the model with imputed missing data is much better than the model without the imputations for missing data. Furthermore, the intercept, CDR, and MMSE are more significant in the new model with imputed values. Also, the plots for the second model are much better in terms of normality, residuals and fit.

Now, we want to simulate a scenario where our data has missing completely at random values. For this, we will take only about 80% of our data which has 348 rows instead of 436. Next, we ran the regression once more and these were the results.

```
adjusteddata<-data[sample(nrow(data),0.8*nrow(data)),]
nrow(adjusteddata)
```

```
## [1] 348
```

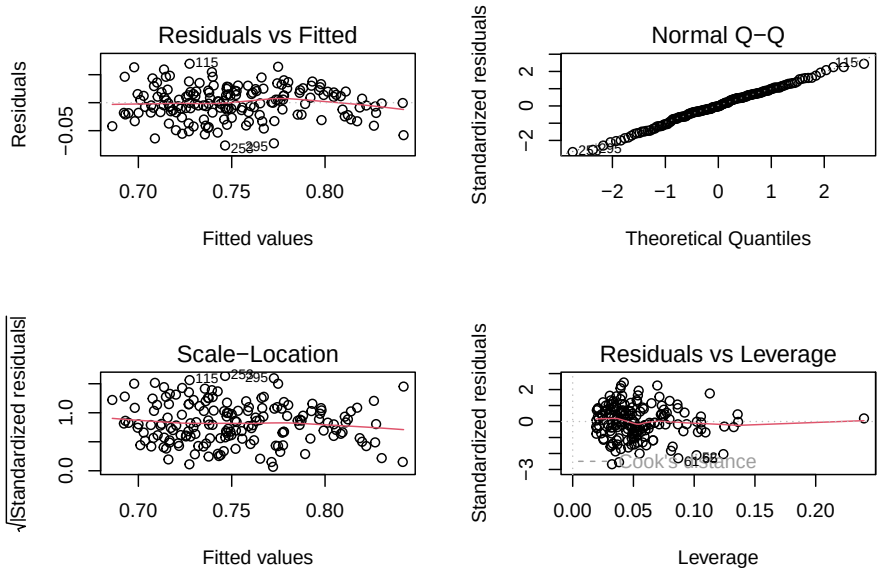
```r
adjustedM.F<-adjusteddata$M.F
adjustednWBV<-adjusteddata$nWBV
adjustedAge<-adjusteddata$Age
adjustedEduc<-adjusteddata$Educ
adjustedSES<-adjusteddata$SES
adjustedMMSE<-adjusteddata$MMSE
adjustedCDR<-adjusteddata$CDR
adjustedeTIV<-adjusteddata$eTIV
adjustedASF<-adjusteddata$ASF

lm(nWBV~M.F+Age+Educ+SES+MMSE+CDR+eTIV+ASF,data = data2)
```

```
##
## Call:
## lm(formula = nWBV ~ M.F + Age + Educ + SES + MMSE + CDR + eTIV +
##     ASF, data = data2)
##
## Coefficients:
## (Intercept)         M.FM          Age         Educ          SES         MMSE
##   8.183e-01   -2.526e-03   -1.890e-03    2.151e-03    2.168e-03    2.717e-03
##         CDR         eTIV          ASF
##  -2.118e-02   -2.627e-05    2.167e-02
```

```r
fit3<-lm(adjustednWBV~adjustedM.F+adjustedAge+adjustedEduc+adjustedSES+adjustedMMSE+adjustedCDR+adjuste
summary(fit3)
```

```
##
## Call:
## lm(formula = adjustednWBV ~ adjustedM.F + adjustedAge + adjustedEduc +
##     adjustedSES + adjustedMMSE + adjustedCDR + adjustedeTIV +
##     adjustedASF, data = adjusteddata)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.076452 -0.017696 -0.000482  0.019628  0.069549
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7.196e-01  2.913e-01   2.471   0.0145 *
## adjustedM.FM -1.105e-02  5.853e-03  -1.888   0.0608 .
## adjustedAge  -2.491e-03  2.022e-04 -12.319   <2e-16 ***
## adjustedEduc -7.373e-04  2.543e-03  -0.290   0.7722
## adjustedSES   2.172e-03  3.069e-03   0.708   0.4802
## adjustedMMSE  1.829e-03  1.065e-03   1.718   0.0878 .
## adjustedCDR  -2.104e-02  9.936e-03  -2.118   0.0357 *
## adjustedeTIV  4.084e-05  9.938e-05   0.411   0.6816
## adjustedASF   8.943e-02  1.208e-01   0.740   0.4601
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02909 on 159 degrees of freedom
##   (180 observations deleted due to missingness)
## Multiple R-squared:  0.6158, Adjusted R-squared:  0.5964
## F-statistic: 31.85 on 8 and 159 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(2,2))
plot(fit3)
```



The model simulated with missing at random values has an equation written as follow:

$$nWBV = 7.474e^{-1} - 7.041e^{-3}x_1 - 2.504e^{-3}x_2 + 7.051e^{-4}x_3 + 2.869e^{-3}x_4 + 1.792e^{-3}x_5 - 2.069e^{-2}x_6 + 2.976e^{-5}x_7 + 7.429e^{-2}x_8$$

This model has quite similar values to the model with missing not at random values. It doesn't change our conclusion at all as we have similar coefficients for each variable and a negligible difference in r^2 fitted value. The missing values are mostly in education and social status, and from our results, these values are insignificant to nWBV. However, to eliminate the bias in our data, we can eliminate those patients who refused to answer the questions fully and therefore we can maximize our data collection.

References

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 45(3), 1–67. https://doi.org/10.18637/jss.v045.i03

https://www.kaggle.com/jboysen/mri-and-alzheimers?select=oasis_longitudinal.csv

https://www.oasis-brains.org/#data